

# Natural Language for Communication (con't.)

Chapter 23.4

# The Machine Translation Problem

মানব পরিবারের সকল সদস্যের সমান ও অবিচ্ছেদ্য অধিকারসমূহ  
এবং সহজাত মর্যাদার স্বীকৃতিই হচ্ছে বিশ্বে শান্তি, স্বাধীনতা এবং  
ন্যায়বিচারের ভিত্তি



**Whereas recognition of the inherent dignity and of the equal and inalienable rights of all members of the human family is the foundation of freedom, justice and peace in the world**

# Brief history

- War-time use of computers in code breaking
- Warren Weaver's memorandum 1949
- Big investment by US Government (mostly on Russian-English)
- Early promise of FAHQT
  - Fully automatic high quality translation

# 1955-1966

- Difficulties soon recognised:
  - no formal linguistics
  - crude computers
  - need for "real-world knowledge"
  - Bar Hillel's "semantic barrier"
- 1966 ALPAC (Automatic Language Processing Advisory Committee) report
  - "insufficient demand for translation"
  - "MT is more expensive, slower and less accurate"
  - "no immediate or future prospect"
  - should invest instead in fundamental computational linguistics research
  - Result: no public funding for MT research in US for the next 25 years (though some privately funded research continued)

# 1966-1985

- Research confined to Europe and Canada
- "2nd generation approach": linguistically and computationally more sophisticated
- c. 1976: success of *Météo* (Canada weather bulletin translation)
- 1978: EC starts discussions of its own MT project, Eurotra
- first commercial systems early 1980s
- FAHQT (fully automatic high quality translation) abandoned in favour of
  - "Translator's Workstation"
  - interactive systems
  - sublanguage / controlled input

# 1985-2000

- Lots of research in Europe and Japan in this "linguistic" paradigm
- PC replaces mainframe computers
- more systems marketed
- despite low quality, users claim increased productivity
- general explosion in translation market thanks to international organizations, globalisation of marketplace ("buy in your language, sell in mine")
- renewed funding in US (work on Farsi, Pashto, Arabic, Korean; include speech translation)
- emergence of new research paradigm ("empirical" methods; allows rapid development of new target language)
- growth of WWW, including translation tools

# Present situation

- creditable commercial systems now available
- wide price range, many very cheap
- MT available free on WWW
- widely used for web-page and e-mail translation
- low-quality output acceptable for reading foreign-language web pages
- but still only a small set of languages covered
- speech translation widely researched

# Why is translation hard (for the computer)?

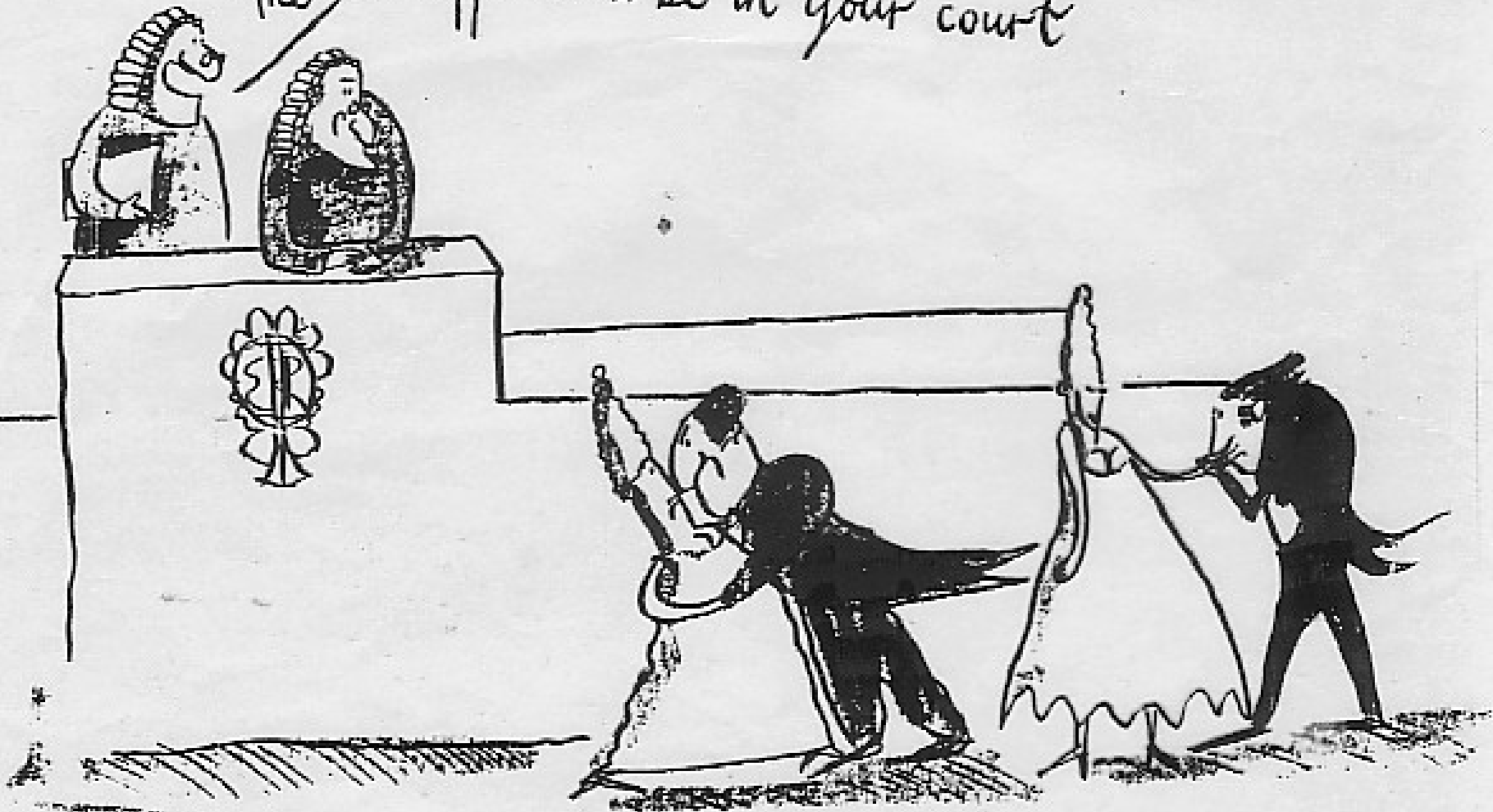
- Two/three steps involved:
  - "Understand" source text
  - Convert that into target language
  - Generate correct target text
- Depends on approach
- Understanding source text involves same problems as for any NLP application



# Understanding the source text

- Lexical ambiguity
  - At morphological level
    - Ambiguity of word vs stem+ending (*tower, flower*)
    - Inflections are ambiguous (*books, loaded*)
    - Derived form may be lexicalised (*meeting, revolver*)
  - Grammatical category ambiguity (eg, *round*)
  - Homonymy
    - Alternate meanings within same grammatical category
    - May or may not be historically or metaphorically related
- Syntactic ambiguity
  - (deep) Due to combination of grammatically ambiguous words
    - *Time flies like an arrow, fruit flies like a banana*
  - (shallow) Due to alternative interpretations of structure
    - *The man saw the girl with a telescope*

The ball appears to be in your court



# Lexical translation problems

- Even assuming monolingual disambiguation ...
- Style/register differences (eg *domicile, merde, medical~anatomical~familiar*)
- Proper names (eg *Addition Barrières*)
- Conceptual differences
- Lexical gaps

# Conceptual differences

- ‘wall’    **German**    *Wand ~ Mauer*
- ‘corner’    **Spanish**    *esquina ~ rincón*
- ‘leg’    **French**    *jambe ~ patte ~ pied*
- ‘leg’    **Spanish**    *pierna ~ pata ~ pie*
- ‘blue’    **Russian**    *голубой ~ синий*
- **Fr.** *louer*    *hire ~ rent*
- **Sp.** *paloma*    *pigeon ~ dove*

- **'rice'**            Malay
  - padi* (harvested grain)
  - beras* (uncooked)
  - nasi* (cooked)
  - emping* (mashed)
  - pulut* (glutinous)
  - bubor* (porridge)
  
- **'wear' ~ 'put on'**    Japanese
  - 羽織る *haoru* (coat, jacket)
  - 穿く *haku* (shoes, trousers)
  - 被る *kaburu* (hat)
  - はめる *hameru* (ring, gloves)
  - 締める *shimeru* (tie, belt, scarf)
  - 付ける *tsukeru* (brooch)
  - 掛ける *kakeru* (glasses)

## Eskimo



- How many words for 'snow' in Eskimo (Inuit)?
- Depending on how you count, between 2 and 12
- About the same as in English!

# Structural translation problems

- Again, even assuming source language disambiguation (though in fact sometimes you might get away with a free ride, esp with "shallow" ambiguities)
- Target language doesn't use the same structure
- Or (worse) it can, but this adds a nuance of meaning

# Structural differences

- adverb → verb
  - Fr. They have just arrived *Ils viennent d'arriver*
  - Sp. We usually go to the cinema *Solemos ir al cine*
  - Ge. I like swimming *Ich schwimme gern*
- adverb → clause
  - Fr. They will probably leave *Il est probable qu'ils partiront*
- Combination can cause problems
  - Fr. They have probably just left
  - \* *Il vient d'être probable qu'ils partent*
  - *Il est probable qu'ils viennent de partir*

# Structural differences

- verb/adverb in Romance languages

Verbs of movement:

Eng. verb expresses manner, adverb  
expresses direction, e.g.

He swam across the river *Il traversa la rivière à la nage*

He rode into town *Il entra en ville à cheval*

We drove from London *Nous venons de Londres en voiture*

The horseman rode into town *Le cavalier entra en ville (à cheval)*

*Un oiseau entra dans la chambre* A bird flew into the room

*Un oiseau entra dans la chambre en sautillant*

\* A bird flew into the room hopping



# Construction is used differently

- Many languages have a "passive" but ...
  - Alternative construction favoured  
These cakes are sold quickly *Ces gâteaux se vendent vite*  
English is spoken here *Ici on parle anglais*
  - Passive may not be available  
Mary was given a book \* *Marie fut donné un livre*  
This bed has been slept in \* *Ce lit a été dormi dans*
  - Passive may be more widely available  
*Ge. Es wurde getanzt und gelacht* There was dancing and laughing  
*Jap. 雨に降られた Ame ni furareta* 'We were fallen by rain'

# Level shift

- Similar grammatical meanings conveyed by different devices

- e.g. definiteness

Da. *hus* 'house' *huset* 'the house' (morphology)

English *the, a, an* etc. (function word)

Rus. *Женщина вышла из дому* ~ *Из дому вышла женщина* (word order)

Jap. どう駅まで行くか (lit. how to station go?)

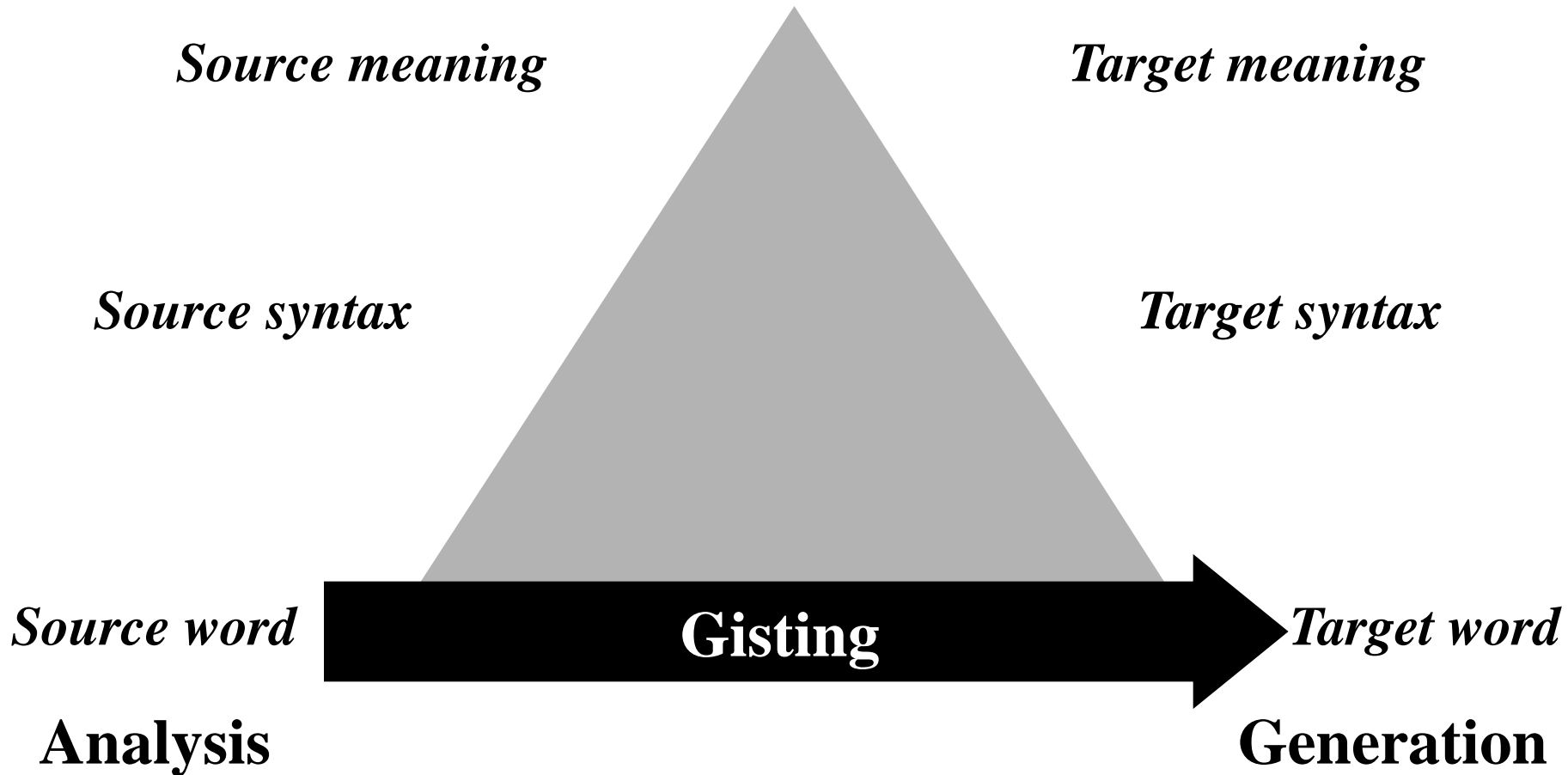
'How do I get to a/the station? (context)

# What's this mean?

- Some of these are difficult problems also for human translators.
- Many require real-world knowledge, intuitions about the meaning of the text, etc. to get a good translation.
- Existing MT systems opt for a strategy of structure-preservation where possible, and do what they can to get lexical choices right.
- First reaction may be that they are rubbish, but when you realise how hard the problem is, you might change your mind.

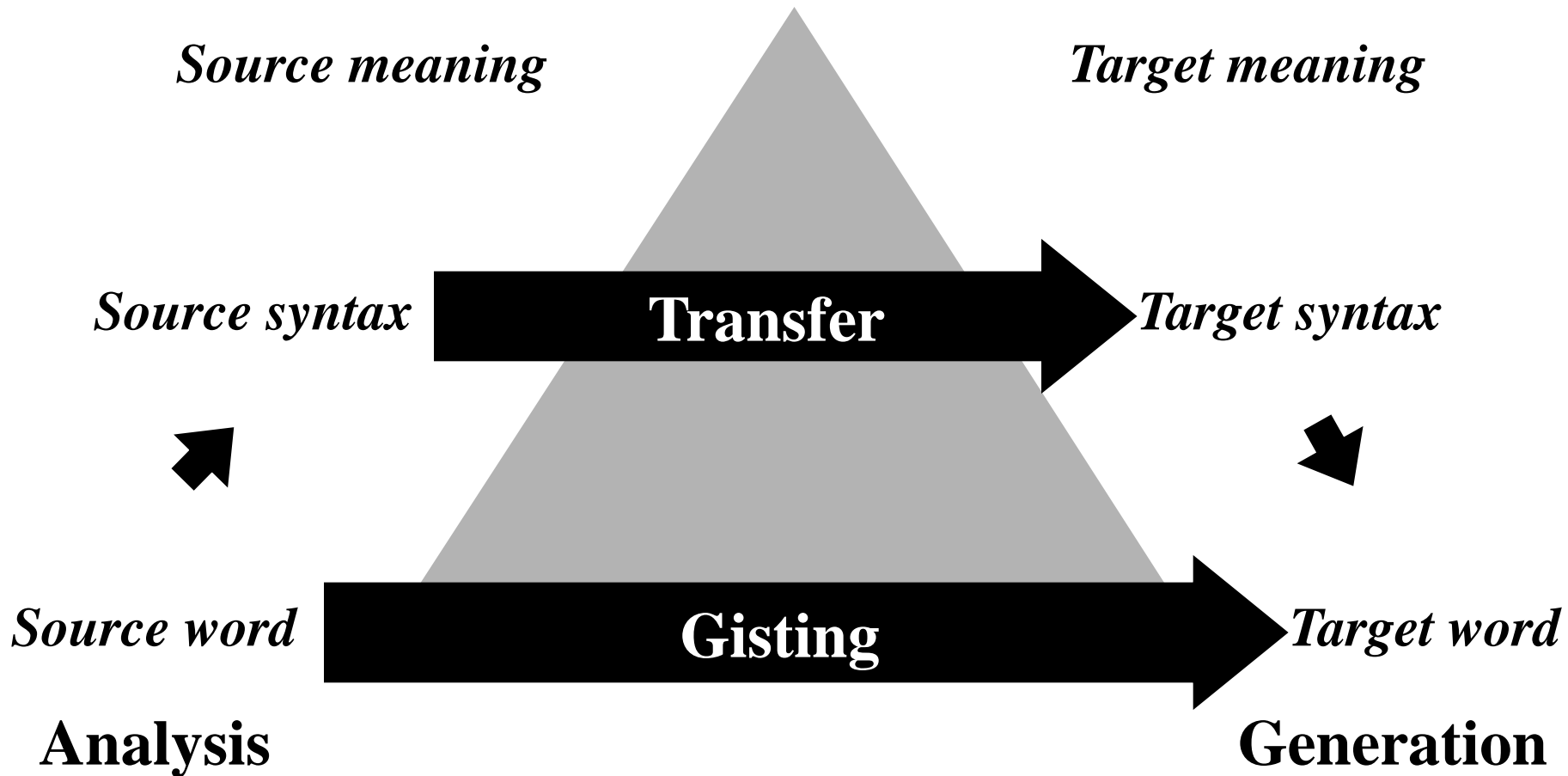
# MT Approaches

## MT Pyramid



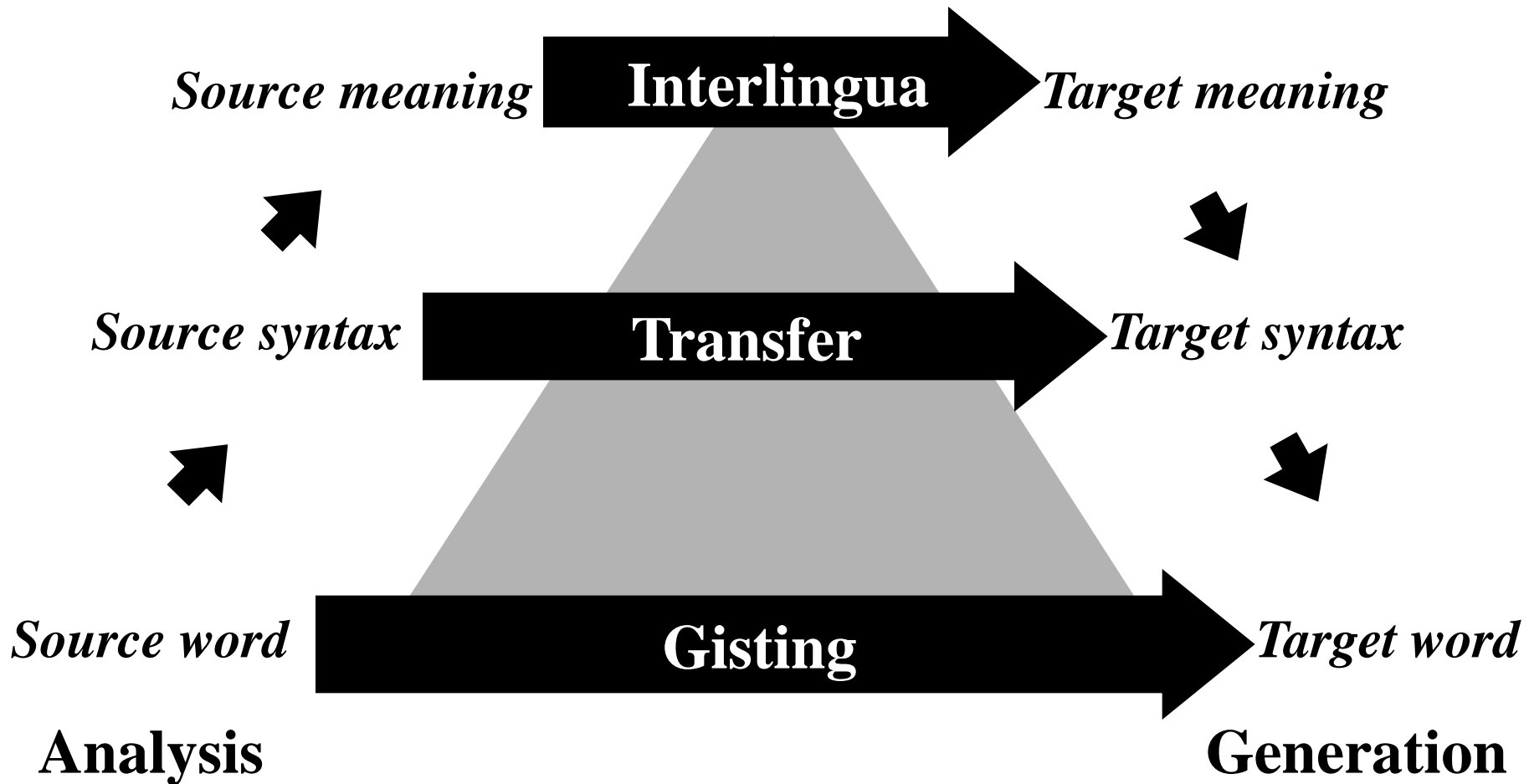
# MT Approaches

## MT Pyramid



# MT Approaches

## MT Pyramid



# Rule-based vs. Data-driven Approaches to MT

- What are the pieces of translation? Where do they come from?
  - **Rule-based:** large-scale "clean" word translation lexicons, manually constructed over time by experts
  - **Data-driven:** broad-coverage word and multi-word translation lexicons, learned automatically from available sentence-parallel corpora
- How does MT put these pieces together?
  - **Rule-based:** large collections of rules, manually developed over time by human experts, that map structures from the source to the target language
  - **Data-driven:** a computer algorithm that explores millions of possible ways of putting the small pieces together, looking for the translation that statistically looks best

# Rule-based vs. Data-driven Approaches to MT

- How does the MT system pick the correct (or best) translation among many options?
  - **Rule-based:** Human experts encode preferences among the rules designed to prefer creation of better translations
  - **Data-driven:** a variety of fitness and preference scores, many of which can be learned from available training data, are used to model a total score for each of the millions of possible translation candidates; algorithm then selects and outputs the best scoring translation



# Rule-based vs. Data-driven Approaches to MT

- Why have the data-driven approaches become so popular?
  - We can now do this!
    - Increasing amounts of sentence-parallel data are constantly being created on the web
    - Advances in machine learning algorithms
    - Computational power of today's computers can train systems on these massive amounts of data and can perform these massive search-based translation computations when translating new texts
  - Building and maintaining rule-based systems is too difficult, expensive and time-consuming
  - In many scenarios, it actually works better!

# Statistical MT (SMT)

- Data-driven, most dominant approach in current MT research
- Proposed by IBM in early 1990s: a direct, purely statistical, model for MT
- Evolved from word-level translation to phrase-based translation
- Main Ideas:
  - **Training:** statistical "models" of word and phrase translation equivalence are learned automatically from bilingual parallel sentences, creating a bilingual "database" of translations
  - **Decoding:** new sentences are translated by a program (the decoder), which matches the source words and phrases with the database of translations, and searches the "space" of all possible translation combinations.

# Statistical MT (SMT)

- Main steps in training phrase-based statistical MT:
  - Create a sentence-aligned parallel corpus
  - **Word Alignment**: train word-level alignment models (*GIZA++*)
  - **Phrase Extraction**: extract phrase-to-phrase translation correspondences using heuristics (*Moses*)
  - **Minimum Error Rate Training (MERT)**: optimize translation system parameters on development data to achieve best translation performance
- Attractive: completely automatic, no manual rules, much reduced manual labor
- Main drawbacks:
  - Translation accuracy levels vary widely
  - Effective only with large volumes (several mega-words) of parallel text
  - Broad domain, but domain-sensitive
  - Viable only for limited number of language pairs!
- Impressive progress in last 5-10 years!

# Statistical MT: Major Challenges

- **Current approaches are too naïve and “direct”:**
  - Good at learning word-to-word and phrase-to-phrase correspondences from data
  - Not good enough at learning how to combine these pieces and reorder them properly during translation
  - Learning general rules requires much more complicated algorithms and computer processing of the data
  - The space of translations that is “searched” often doesn’t contain a perfect translation
  - The fitness scores that are used aren’t good enough to always assign better scores to the better translations → we don’t always find the best translation even when it’s there!
  - MERT is brittle, problematic and metric-dependent!
- **Solutions:**
  - Google solution: more and more data!
  - Research solution: “smarter” algorithms and learning methods

# Statistical MT Systems

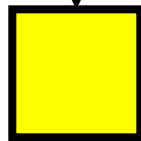
Spanish/English  
Bilingual Text

English  
Text

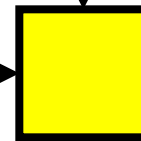
Statistical Analysis

Statistical Analysis

Spanish



Broken  
English



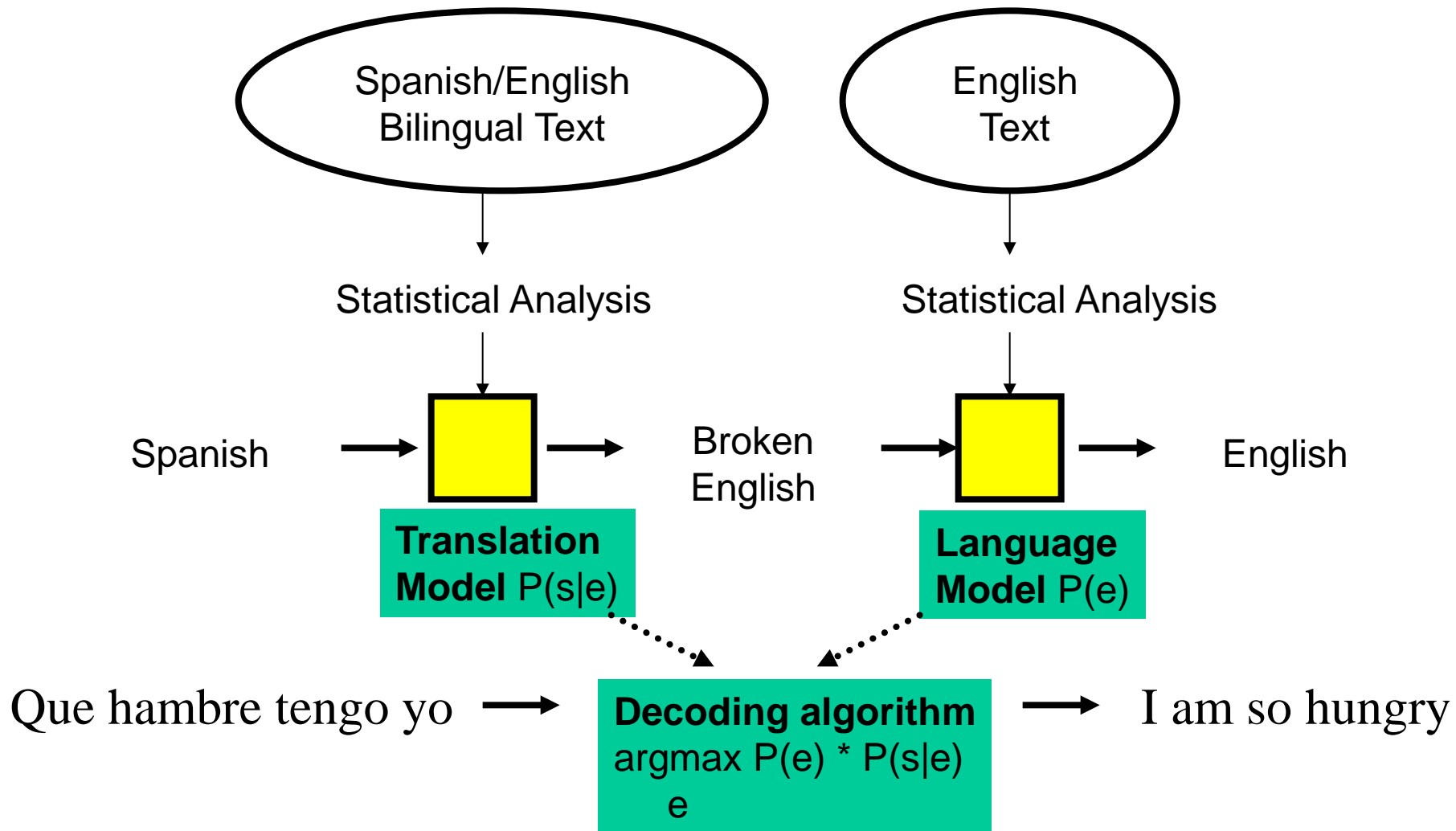
English

Que hambre tengo yo

What hunger have I,  
Hungry I am so,  
I am so hungry,  
Have I that hunger ...

I am so hungry

# Statistical MT Systems



# Translation and Alignment

Translations are expensive to commission

Generally SMT research relies on already existing translations

- These typically come in the form of aligned documents.

A sentence alignment, using pre-existing document boundaries, is performed automatically.

- Low-scoring or non-one-to-one sentence alignments are discarded.
- The resulting aligned sentences constitute the training data.

# Target Language Models

The translation problem can be described as modeling the probability distribution  $P(E|F)$ , where  $F$  is a string in the source language and  $E$  is a string in the target language.

Using Bayes' Rule, this can be rewritten

$$P(E|F) = \frac{P(F|E)P(E)}{P(F)}$$

$$= P(F|E)P(E) \quad [\text{since } F \text{ is observed as the sentence to be translated, } P(F)=1]$$

$P(F|E)$  is called the "translation model" (TM).

$P(E)$  is called the "language model" (LM).

The LM should assign probability to sentences which are "good English".



# Target Language Models

- Typically, N-Gram language models are employed
- These are finite state models which predict the next word of a sentence given the previous several words. The most common N-Gram model is the trigram, wherein the next word is predicted based on the previous 2 words.
- The job of the LM is to take the possible next words that are proposed by the TM, and assign a probability reflecting whether or not such words constitute "good English".

---

$p(\text{the}|\text{went to})$

$p(\text{happy}|\text{was feeling})$

$p(\text{time}|\text{at the})$

$p(\text{the}|\text{took the})$

$p(\text{sagacious}|\text{was feeling})$

$p(\text{time}|\text{on the})$

# Resource Availability

Most statistical machine translation (SMT) research has focused on a few "high-resource" Languages (European, Chinese, Japanese, Arabic).

Some other work: translation for the rest of the world's languages found on the web.

# Resource Availability

Approximate  
Parallel Text Available  
(with English)

(~200M words)

Various  
Western European  
languages:  
parliamentary  
proceedings,  
govt documents  
(~30M words)

Bible/Koran/  
Book of Mormon/  
Dianetics  
(~1M words)

Nothing/  
Univ. Decl.  
Of Human  
Rights  
(~1K words)



Chinese French Arabic Italian Danish Finnish Serbian Bengali Uzbek Chechen Khmer

# Four Problems for Statistical MT

- Language model
  - Given an English string  $e$ , assigns  $P(e)$  by the usual methods we've been using sequence modeling.
- Translation model
  - Given a pair of strings  $\langle f, e \rangle$ , assigns  $P(f | e)$  again by making the usual Markov assumptions
- Training
  - Getting the numbers needed for the models
- Decoding algorithm
  - Given a language model, a translation model, and a new sentence  $f$  ... find translation  $e$  maximizing  $P(e) * P(f | e)$

# Language Model Trivia

- *Google Ngrams data*

- Number of tokens: 1,024,908,267,229
- Number of sentences: 95,119,665,584
- Number of unigrams: 13,588,391
- Number of bigrams: 314,843,401
- Number of trigrams: 977,069,902
- Number of four grams: 1,313,818,354
- Number of five grams: 1,176,470,663

# Alignment Probabilities

- Recall what all of the models are doing

$$\text{Argmax } P(e|f) = P(f|e)P(e)$$

In the simplest models  $P(f|e)$  is just direct word-to-word translation probs. So let's start with how to get those, since they're used directly or indirectly in all the models.

# Training alignment probabilities

- Step 1: Get a parallel corpus
  - Hansards
    - Canadian parliamentary proceedings, in French and English
    - Hong Kong Hansards: English and Chinese
- Step 2: Align sentences
- Step 3: Use EM to train word alignments. Word alignments give us the counts we need for the word to word  $P(f|e)$  probs

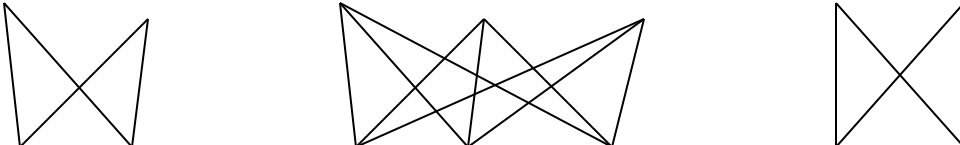
# Step 3: Word Alignments

- Of course, sentence alignments aren't what we need. We need word alignments to get the stats we need.
- It turns out we can bootstrap word alignments from raw sentence aligned data (no dictionaries)
- Using EM
  - Recall the basic idea of EM. A model predicts the way the world should look. We have raw data about how the world looks. Start somewhere and adjust the numbers so that the model is doing a better job of predicting how the world looks.



# EM Training: Word Alignment Probs

... la maison ... la maison bleue ... la fleur ...  
... the house ... the blue house ... the flower ...



All word alignments equally likely

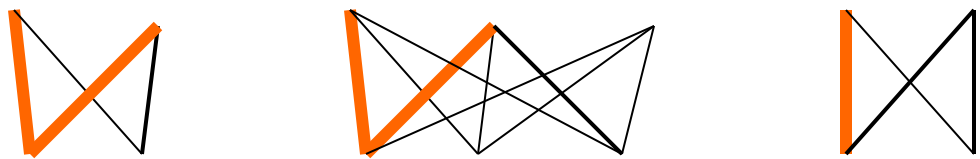
All  $P(\text{french-word} \mid \text{english-word})$  equally likely.

# EM Training Constraint

- Recall what we're doing here... Each English word has to translate to some French word.
- But its still true that

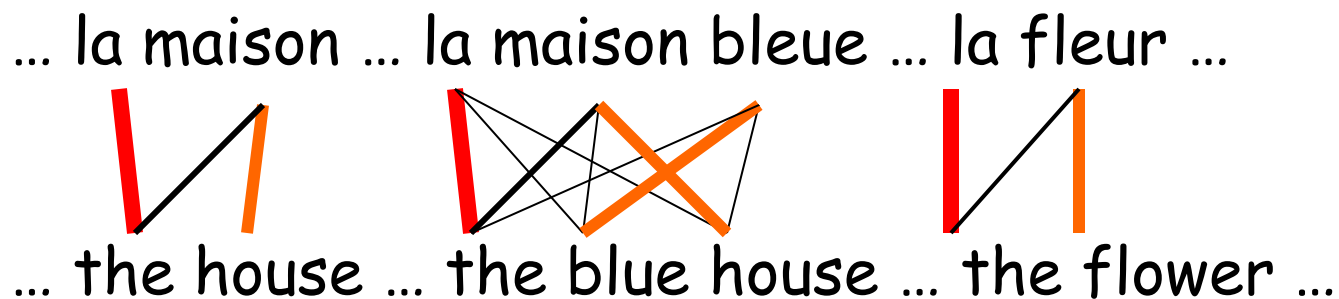
$$\sum_i P(f_i|E) = 1$$

# EM for training alignment probs

... la maison ... la maison bleue ... la fleur ...  
  
... the house ... the blue house ... the flower ...

"la" and "the" observed to co-occur frequently,  
so  $P(\text{la} \mid \text{the})$  is increased.

# EM for training alignment probs

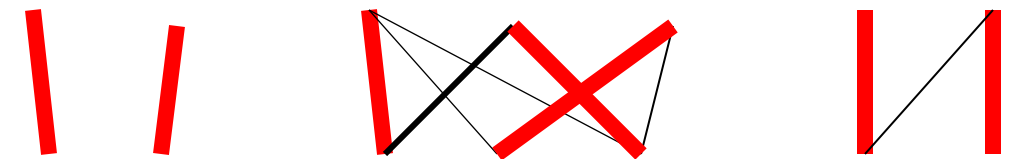


"house" co-occurs with both "la" and "maison", but  $P(\text{maison} \mid \text{house})$  can be raised without limit, to 1.0, while  $P(\text{la} \mid \text{house})$  is limited because of "the"

(pigeonhole principle)

# EM for training alignment probs

... la maison ... la maison bleue ... la fleur ...  
... the house ... the blue house ... the flower ...




The diagram illustrates the EM algorithm for training word alignment probabilities. It shows two rows of text: the top row in French and the bottom row in English. Red vertical bars are placed under the words in both rows to indicate alignment. In the middle, a black line connects 'maison' to 'house' and 'bleue' to 'house', with a red 'X' over the crossing lines, indicating a correction from a previous iteration.

settling down after another iteration

# EM for training alignment probs

... la maison ... la maison bleue ... la fleur ...  
... the house ... the blue house ... the flower ...



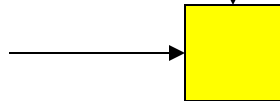
**Inherent hidden structure revealed by EM training!**

# Direct Translation

... la maison ... la maison bleue ... la fleur ...  
... the house ... the blue house ... the flower ...

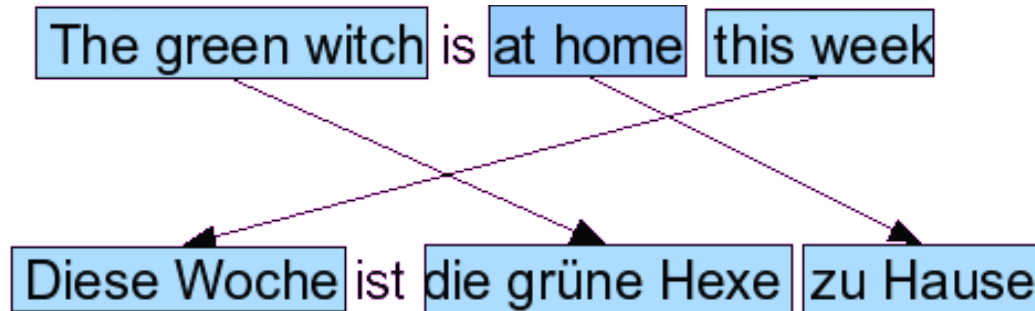
$P(\text{juste} \mid \text{fair}) = 0.411$   
 $P(\text{juste} \mid \text{correct}) = 0.027$   
 $P(\text{juste} \mid \text{right}) = 0.020$   
...

New  
French  
sentence



Possible English translations,  
rescored by language model

# Phrase-Based Translation



- *Generative story here has three steps*
  - 1) Discover and align phrases during training
  - 2) Align and translate phrases during decoding
  - 3) Finally move the phrases around



# Phrase-based MT

- Language model  $P(E)$
- Translation model  $P(F|E)$ 
  - Model
  - How to train the model
- Decoder: finding the sentence  $E$  that is most probable

# Generative story again

- 1) Group English source words into phrases  $e_1, e_2, \dots, e_n$
- 2) Translate each English phrase  $e_i$  into a Spanish phrase  $f_j$ .
  - The probability of doing this is  $\phi(f_j|e_i)$
- 3) Then (optionally) reorder each Spanish phrase
  - We do this with a **distortion** probability
  - A measure of distance between positions of a corresponding phrase in the 2 languages
  - "What is the probability that a phrase in position X in the English sentences moves to position Y in the Spanish sentence?"

# Distortion probability

- The distortion probability is parameterized by:
  - The start position of the foreign (Spanish) phrase generated by the  $i$ th English phrase  $e_i$ .
  - The end position of the foreign (Spanish) phrase generated by the  $i-1$ th English phrase  $e_{i-1}$ .
- We'll call the distortion probability  $d(\cdot)$

# Final translation model for phrase-based MT

$$P(f, d | e) = \prod_i P(f_i | e_i) P(d_i)$$

<b>Position</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>
<b>English</b>	Mary	did not	slap	the	green witch
<b>Spanish</b>	Maria	no	dió una bofetada	a la	bruja verde

$$\begin{aligned}
 = & P(\text{Maria, Mary}) \times d(1) \times P(\text{no|did not}) \times d(1) \times \\
 & P(\text{dió una bofetada|slap}) \times d(1) \times P(\text{a la|the}) \times d(1) \times \\
 & P(\text{bruja verde|green witch}) \times d(1)
 \end{aligned}$$

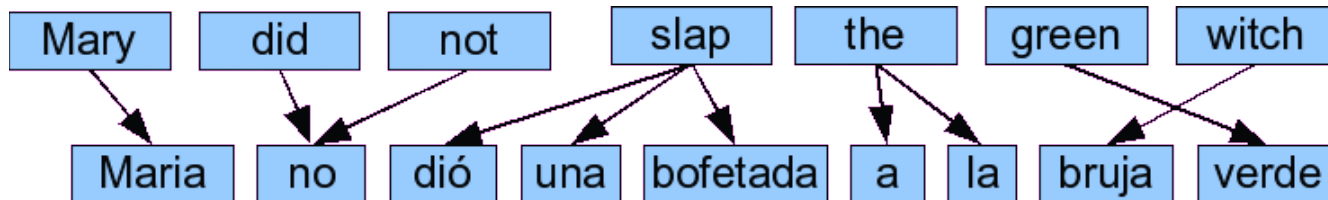
# Training $P(F|E)$

- What we mainly need to train is  $\phi(f_j|e_i)$
- Assume as before we have a large bilingual training corpus
- And suppose we knew exactly which phrase in Spanish was the translation of which phrase in the English
- We call this a **phrase alignment**
- If we had this, we could just count-and-divide:

$$\phi(\bar{f}, \bar{e}) = \frac{\text{count}(\bar{f}, \bar{e})}{\sum_{\bar{f}} \text{count}(\bar{f}, \bar{e})}$$

# But we don't have phrase alignments

- What we have instead are word alignments:





# Final Problem

- Decoding...
  - Given a trained model and a foreign sentence produce
    - $\text{Argmax } P(e|f)$
    - Can't use Viterbi it's too restrictive
    - Need a reasonable efficient search technique that explores the sequence space based on how good the options look...
      - $A^*$



$A^*$

- Recall for  $A^*$  we need
  - Goal State
  - Operators
  - Heuristic

# A\*

- Recall for A\* we need
  - Goal State      Good coverage of **source**
  - Operators      Translation of phrases/words  
distortions  
deletions/insertions
  - Heuristic      Probabilities (tweaked)

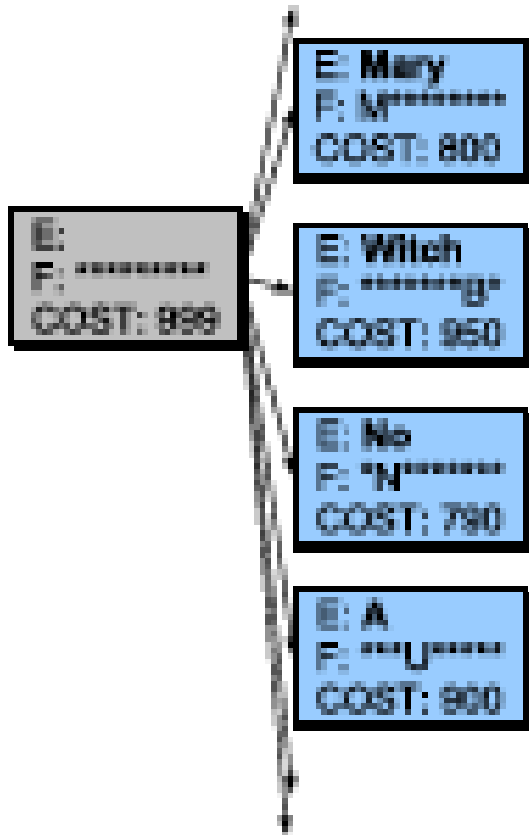
# A\* Decoding

- Why not just use the probability as we go along?
  - Turns it into Uniform-cost not A\*
  - That favors shorter sequences over longer ones.
  - Need to counter-balance the probability of the translation so far with its "progress towards the goal".

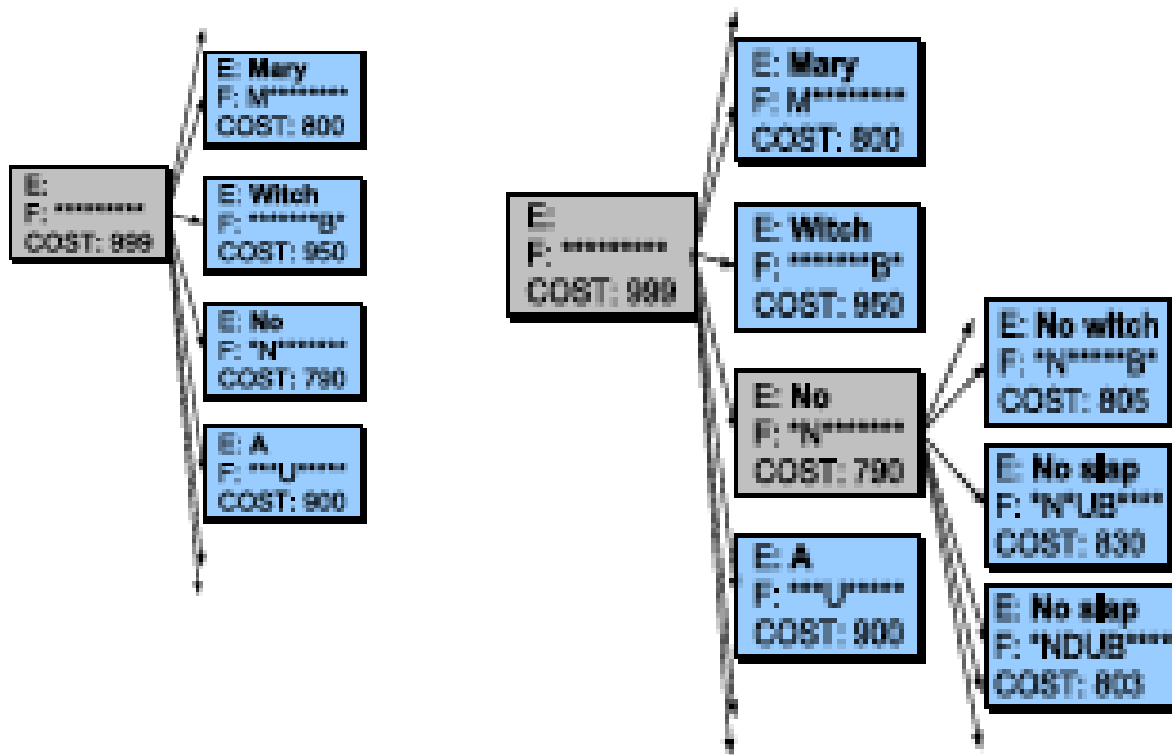
# A\*/Beam

- Sorry...
  - Even that doesn't work because the space is too large
  - So as we go we'll prune the space as paths fall below some threshold

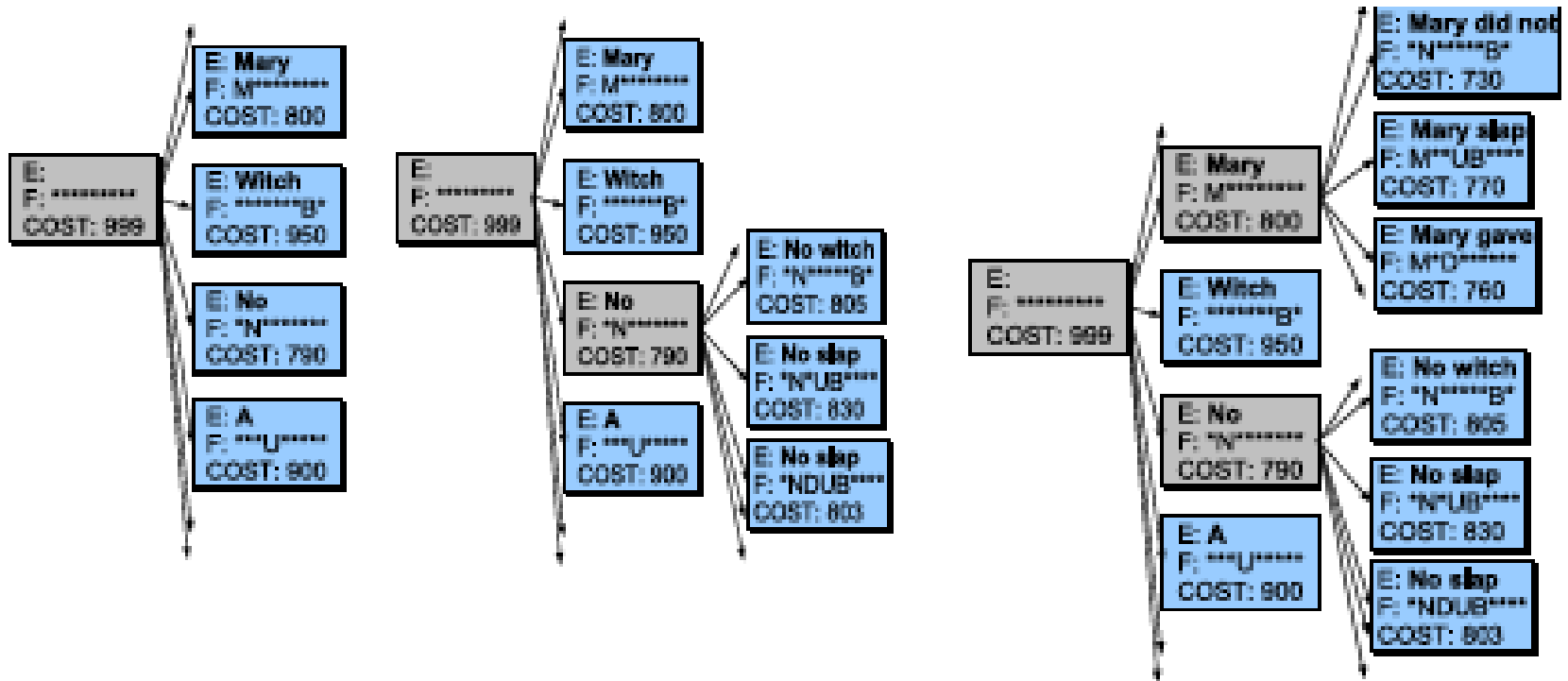
# A\* Decoding



# A\* Decoding



# A\* Decoding



# Evaluation

- There are 2 dimensions along which MT systems can be evaluated
  - Fluency
    - How good is the output text as an example of the target language
  - Fidelity
    - How well does the output text convey the source text
      - Information content and style



# How to Evaluate MT Results?

Compare current translation to:

- Idea #1: a human translation. OK, but:
  - Good translations can be very dissimilar
  - We'd need to find hidden features (e.g. alignments)
- Idea #2: other top  $n$  translations (the "n-best list"). Better in practice, but
  - Many entries in n-best list are the same apart from hidden links
- Compare with a **loss function  $\mathcal{L}$** 
  - 0/1: wrong or right; equal to reference or not
  - Task-specific metrics (word error rate, BLEU, ...)

# Evaluating MT: Human tests for fluency

- Rating tests: Give the raters a scale (1 to 5) and ask them to rate
  - Or distinct scales for
    - Clarity, Naturalness, Style
  - Or check for specific problems
    - Cohesion (Lexical chains, anaphora, ellipsis)
      - Hand-checking for cohesion.
    - Well-formedness
      - 5-point scale of syntactic correctness

# Evaluating MT: Human tests for fidelity

- Adequacy
  - Does it convey the information in the original?
  - Ask raters to rate on a scale
    - Bilingual raters: give them source and target sentence, ask how much information is preserved
    - Monolingual raters: give them target + a good human translation

# Evaluating MT: Human tests for fidelity

- Informativeness
  - Task based: is there enough info to do some task?

# Human Evaluation

**Je suis fatigué.**

**Tired is I.**

**Cookies taste good!**

**I am exhausted.**

<b>Adequacy</b>	<b>Fluency</b>
<b>5</b>	<b>2</b>
<b>1</b>	<b>5</b>
<b>5</b>	<b>5</b>

# Human Evaluation

PRO

High quality

CON

Expensive!

Person (preferably bilingual) must make a time-consuming judgment per system hypothesis.

Expense prohibits frequent evaluation of incremental system modifications.

# Automatic Evaluation

## PRO

Cheap. Given available reference translations, free thereafter.

## CON

We can only measure some proxy for translation quality. (Such as N-Gram overlap or edit distance).

# BiLingual Evaluation Understudy (BLEU)

- Automatic Technique
- Requires the pre-existence of Human (Reference) Translations
- Approach:
  - Produce corpus of high-quality human translations
  - Judge "closeness" numerically (word-error rate)
  - Compare n-gram matches between candidate translation and 1 or more reference translations




# Automatic Evaluation: Bleu Score

*N-Gram  
precision*

$$p_n = \frac{\sum_{n\text{-gram} \in \text{hyp}} \text{count}_{clip}(n\text{-gram})}{\sum_{n\text{-gram} \in \text{hyp}} \text{count}(n\text{-gram})}$$

*Bounded above  
by highest count  
of n-gram in any  
reference sentence*



*brevity  
penalty*

$$B = \begin{cases} e^{(1 - |\text{ref}| / |\text{hyp}|)} & \text{if } |\text{ref}| > |\text{hyp}| \\ 1 & \text{otherwise} \end{cases}$$

*Bleu score:  
brevity penalty,  
geometric  
mean of N-Gram  
precisions*

$$\text{Bleu} = B \cdot \exp \left[ \frac{1}{N} \sum_{n=1}^N p_n \right]$$

# BLEU Evaluation Metric

## Reference (human) translation:

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

## Machine translation:

The American [?] international airport and its the office al receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

- N-gram precision (score is between 0 & 1)
    - What percentage of machine n-grams can be found in the reference translation?
      - An n-gram is an sequence of n words
      - Not allowed to use same portion of reference translation twice (can't cheat by typing out "the the the the the")
  - Brevity penalty
    - Can't just type out single word "the" (precision 1.0!)
- \*\*\* Amazingly hard to "game" the system (i.e., find a way to change machine output so that BLEU goes up, but quality doesn't)

# BLEU Evaluation Metric

## Reference (human) translation:

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

## Machine translation:

The American [?] international airport and its the office al receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

- BLEU4 formula  
(counts n-grams up to length 4)

$$\exp (1.0 * \log p1 + 0.5 * \log p2 + 0.25 * \log p3 + 0.125 * \log p4 - \max(\text{words-in-reference} / \text{words-in-machine} - 1, 0))$$

p1 = 1-gram precision

P2 = 2-gram precision

P3 = 3-gram precision

P4 = 4-gram precision

# Multiple Reference Translations

## Reference translation 1:

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

## Reference translation 2:

Guam International Airport and its offices are maintaining a high state of alert after receiving an e-mail that was from a person claiming to be the wealthy Saudi Arabian businessman Bin Laden and that threatened to launch a biological and chemical attack on the airport and other public places .

## Machine translation:

The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

## Reference translation 3:

The US International Airport of Guam and its office has received an email from a self-claimed Arabian millionaire named Laden , which threatens to launch a biochemical attack on such public places as airport . Guam authority has been on alert .

## Reference translation 4:

US Guam International Airport and its office received an email from Mr. Bin Laden and other rich businessman from Saudi Arabia . They said there would be biochemistry air raid to Guam Airport and other public places . Guam needs to be in high precaution about this matter .

# Bleu Comparison

## Chinese-English Translation Example:

**Candidate 1:** It is a guide to action which ensures that the military always obeys the commands of the party.

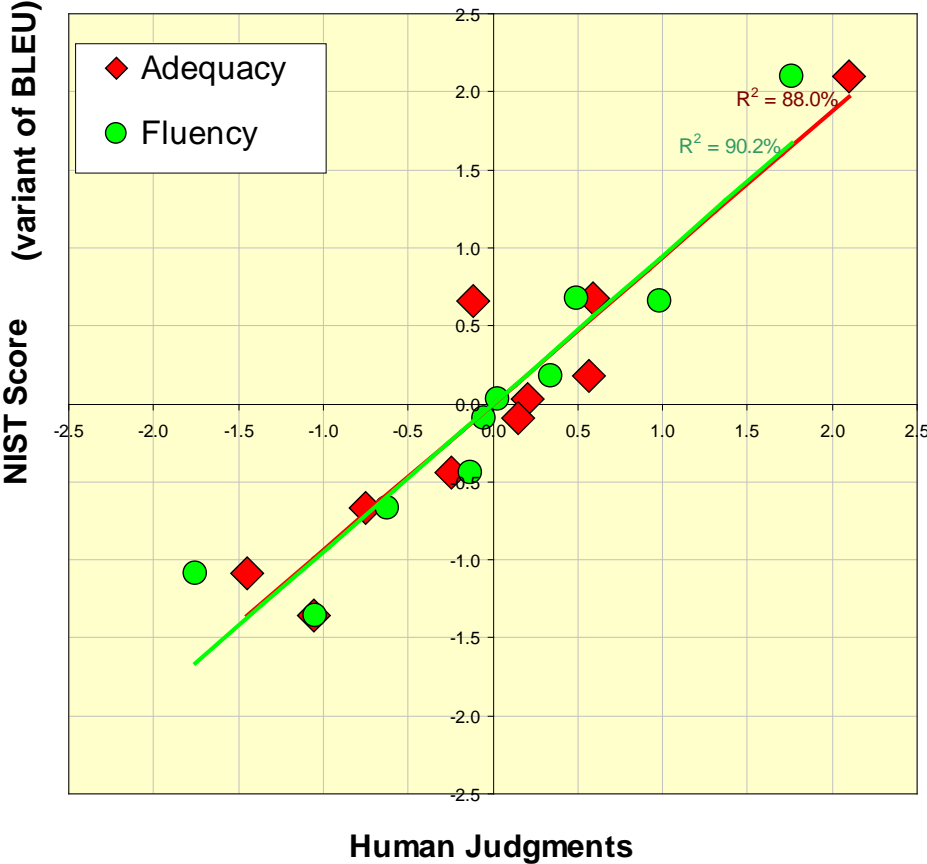
**Candidate 2:** It is to insure the troops forever hearing the activity guidebook that party direct.

**Reference 1:** It is a guide to action that ensures that the military will forever heed Party commands.

**Reference 2:** It is the guiding principle which guarantees the military forces always being under the command of the Party.

**Reference 3:** It is the practical guide for the army always to heed the directions of the party.

# BLEU Tends to Predict Human Judgments



# Summary of MT

- Lots of machine translation systems have been implemented
- Statistical methods based on phrase frequencies are currently most successful